EM Algorithm

1. Introduction

1. Necessity of EM algorithm

표본 데이터의 확률 분포를 추정하는 경우를 생각해보자. 특정 분포를 가정한다고 한다면, 확률 분포를 추정하는 문제는 가정한 분포의 모수를 추정하는 것이 된다. 우리는 이 경우에 관측된(observed) 데이터의 결합확률분포를 최대화하는 모수를 추정하게 된다. 이 방법이 바로 Maximum likelihood estimation (MLE) 이다.

우리는 종종 latent (또는 unobserved) 데이터를 포함하여 모수를 추정해야 하는 경우가 있다. 예를 들어, 아래 데이터 분포를 추정하는 문제가 바로 그 경우다.

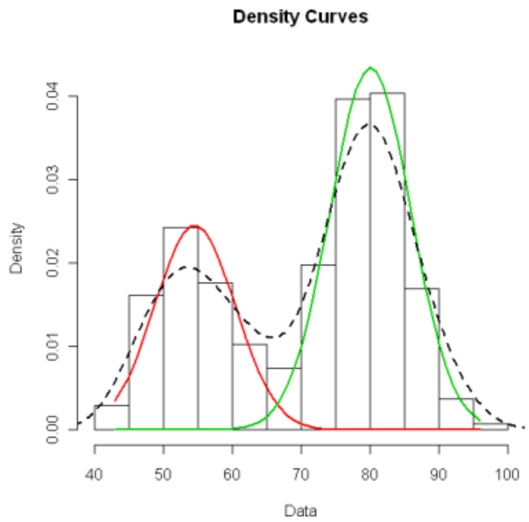


Figure 1: Mixture of Gaussian Problem [1]

원래 데이터는 빨간색과 초록색의 혼합 정규분포로부터 임의로 추출되었지만, 우리는 이 정보를 모른채 검정색 점선의 데이터 분포만으로 모수를 추정한다. 이때 빨간색과 초록색이라는 label이 우리가 모르는 unobserved 데이터가 된다.

당연히 이 label을 알고 모수를 추정하면 쉽겠지만 잠재된 변수이기 때문에 알기 어렵다.

그렇다고 단일 정규분포로 가정하여 모수를 추정하자니, 딱 봐도 (현실적으로는 어렵겠지만) 데이터가 두 개의 혼합 분포로 이루어져 있어 단일 분포 가정으론 추정이 잘 되기 어려울 것 같다.

그래서 모수를 추정할 때 unobserved 데이터를 observed 데이터와 함께 사용하여(정보를 최대한 활용하는 게 좋으니까) Maximum likelihood를 갖는 모수를 추정하게 되는데, 우리가 기존에 알고 있던 MLE 방법으로는 이것이 매우 어렵다. 그래서 모수를 추정하기 위해 새롭게 도입된 방법이 바로 EM 알고리즘이다.

2. Jensen’s inequality

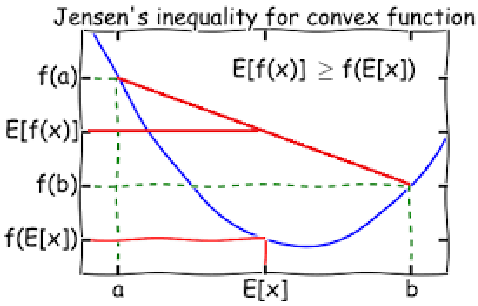
EM 알고리즘에 대해 이해하기 위해서 Jensen의 부등식에 대한 개념이 선행되어야 한다. 추가로 Convex와 Concave에 대한 개념은 다른 글들을 통해 숙지하길 바란다.

Jensen's inequality says

- If function f is convex,

$$E(f(x)) \geq f(E(x))$$

그림을 통해 직관적으로 이해가 가능하다



- If function f is concave,

$$E(f(x)) \leq f(E(x))$$

위 그림의 파란색 그래프가 y축 방향으로 뒤집어진 된 경우라고 생각하면 직관적으로 이해할 수 있다.

2. EM Algorithm

다음과 같은 training set $\{x^{(1)}, \cdots, x^{(m)}\}$ 이 있다고 가정하자. 우리는 training set과 latent variable z를 포함한 모델 $p(x, z; \theta)$의 파라미터 θ를 추정하고자 한다. 그러면 우리는 다음과 같은 log-likelihood function $l(\theta)$를 세울 수 있다.

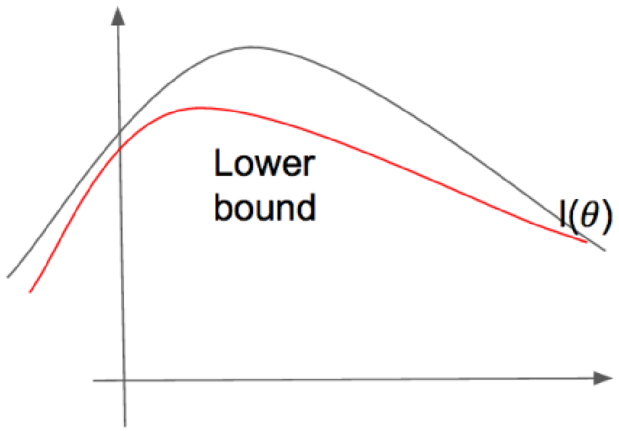
$$\begin{aligned} l(\theta) &= \sum_{i=1}^{m} \log p(x^{(i)}; \theta) \\ &= \sum_{i=1}^{m} \log \sum_{z} p(x^{(i)}, z^{(j)}; \theta) \end{aligned}$$

위 식을 최대화하는 파라미터 값을 찾으면 되겠지만, log안에 $\sum$가 있고, $z^{(j)}$가 latent random variable이기 때문에 때문에 최적화 문제가 쉽지 않다(i.e. difficult non-convex optimization problem).

그래서 파라미터를 찾기 위해 EM 알고리즘이 필요하고 EM 알고리즘은 크게 E-step과 M-step 두 가지로 나뉜다.

- **E(Expectation)-step** : $l(\theta)$의 lower bound를 세운다.
- **M(Maximization)-step) :** lower bound를 최대화한다.

위 두 step을 반복하며 파라미터를 찾는다.



수식으로 이해하기 전, 위 그림을 보면 그 과정을 직관적으로 이해할 수 있다. $l(\theta)$를 직접 최적화하는 것은 어려우니 빨간선인 lower bound를 찾고, 이걸 반복적으로 최대화하면서 $l(\theta)$와의 간격을 좁히는 것이 EM 알고리즘이다.

1. E-step

Case - single training set

가장 먼저 lower bound를 세우는 과정인 E-step이다. Expectation step을 말하는데, lower bound가 latent variable z와 관련된 Expectation의 꼴로 만들어지기 때문에 E-step이라고 한다.

파라미터를 찾기 위해

$$l(\theta) = \sum_{i=1}^{m} \log p(x^{(i)};\theta)$$

를 최대화해야 하는데, 식을 간단하게 쓰고 정리하기 위해 single training set ($m = 1$)인 경우라고 가정하자. 그렇다면 기존 문제는 $\log p(x;\theta)$를 최대화하는 것과 같아진다.

$\log p(x;\theta)$을 직접 최대화하지 못하므로 Jensen 부등식을 이용해 $\log p(x;\theta)$의 lower bound를 세워야 하고, 이를 위해 z에 대한 새로운 distribution $Q(z)$가 필요하다. ($\sum_z Q(z) = 1, Q(z) \geq 0$

※ 만약 z가 continuous하다면, $\int$을 사용한다.

※ $Q(z)$가 필요한 이유는 그냥 식을 도출하기 위한 trick이라고 이해하면 된다.

lower bound를 세우는 과정은 아래와 같다.

$$\begin{aligned}\log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\ &= \log \sum_z Q(z)\frac{p(x,z;\theta)}{Q(z)} \\ &\geq \sum_z Q(z)\log \frac{p(x,z;\theta)}{Q(z)}\end{aligned}$$

첫 번째 줄 → 두 번째 줄은 그냥 등식을 성립하게 하기 위해 분모와 분자에 $Q(z)$를 넣어준 것이고, **두 번째 줄 → 세 번째 줄**은 앞서 설명했던 Jensen의 부등식으로부터 도출된다.

두 번째줄 → 세 번째 줄의 과정을 좀 더 자세히 설명하면 다음과 같다.

1) 기대값의 정의에 의해,

$$E_{z\sim Q}[\frac{p(x,z;\theta)}{Q(z)}] = \sum_z Q(z)\frac{p(x,z;\theta)}{Q(z)}$$

2) $p(x,z;\theta)/Q(z) = K$라고 한다면, log함수는 concave하므로 Jensen의 부등식에 의해

$$\begin{aligned} f(E[K]) \geq E[f(K)], \quad \text{where } f(x) = \log x \\ \Leftrightarrow f(E_{z\sim Q}[\frac{p(x,z;\theta)}{Q(z)}]) \geq E_{z\sim Q}[f(\frac{p(x,z;\theta)}{Q(z)})] \end{aligned}$$

가 성립한다. (좀 더 정확히 말하자면, $Q(z) \neq 0$이어야 한다.)

즉 z의 **분포 $Q(z)$가 어떻든 간에** 아래 부등식이 성립한다.

$$\log p(x;\theta) \geq \sum_z Q(z)\log \frac{p(x,z;\theta)}{Q(z)}$$

여기서 bound를 타이트하게 만들수록 최적화하고자 하는 파라미터를 찾기 쉬워질 것이다. 타이트하게 만들기 위해서는 위 부등식의 등호가 성립하면 된다.

$$f(E_{z\sim Q}[\frac{p(x,z;\theta)}{Q(z)}]) \geq E_{z\sim Q}[f(\frac{p(x,z;\theta)}{Q(z)})]$$

이는 곧 위 Jensen 부등식에서 등호가 성립하는 것과 동치이고, 등호가 성립하기 위해서는 $\frac{p(x,z;\theta)}{Q(z)}$가 z에 독립인 상수가 되야 한다.

이 사실을 통해 새로운 식을 도출할 수 있다.

등호가 성립하기 위해 $p(x,z;\theta)/Q(z) = c$ (constant) 이어야 하고, 이는 곧 $Q(z) \propto p(x,z;\theta)$ 가 성립한다는 것이다.

추가로 $Q(z)$는 확률분포이기에 $\sum_z Q(z) = 1$ 를 만족해야 하므로 $Q(z)$를 $p(x,z;\theta)$에 대한 비율로 표현이 가능해진다. 즉,

$$\begin{aligned} Q(z) &= \frac{p(x,z;\theta)}{\sum_z p(x,z;\theta)} \\ &= \frac{p(x,z;\theta)}{p(x;\theta)} \\ &= p(z|x;\theta) \end{aligned}$$

가 된다.

따라서 우리는 임의의 분포 $Q(z)$에 대해 부등식의 등호가 성립하게 만드는 분포는 x, θ가 주어졌을 때 z의 사후분포와 같다고 여길 수 있다.

위 과정을 쿨백-라이블러 발산의 관점에서도 이해할 수 있는데 구글에 꽤 있으니 궁금한 사람은 참고하길 바란다.

다시 말하자면, lower bound가

$$E_{z\sim Q}[f(\frac{p(x,z;\theta)}{Q(z)})] = \sum_z Q(z)\log \frac{p(x,z;\theta)}{Q(z)}$$

와 같이 기댓값의 꼴이기 때문에 E-step이라고 불린다.

Case - multiple training sets

자 이제 앞서 가정했던 single training set을 m개의 training set으로 확장시켜 lower bound를 세워보자

$$
\begin{aligned}
l(\theta) &= \sum_{i=1}^{m} \log p(x^{(i)};\theta) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(j)}} p(x^{(i)}, z^{(j)};\theta) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(j)}} Q_i(z^{(j)}) \frac{p(x^{(i)}, z^{(j)};\theta)}{Q_i(z^{(j)})} \\
&\geq \sum_{i=1}^{m} \sum_{z^{(j)}} Q_i(z^{(j)}) \log \frac{p(x^{(i)}, z^{(j)};\theta)}{Q_i(z^{(j)})}
\end{aligned}
$$

lower bound를 위와 같이 정의할 수 있고, 동일한 방법으로

$$
Q_i(z^{(j)}) = p(z^{(j)}|x^{(i)};\theta)
$$

를 도출할 수 있다.

2. M-step

다음으로 E-step에서 정의된 lower bound를 최대로 만드는 파라미터를 업데이트 하는 단계인 M-step이다.

앞서 $z^{(j)}$의 사후분포를 구한 식을 이용하면

$$
Q_i(z^{(j)}) = p(z^{(j)}|x^{(i)};\theta)
$$
$$
l(\theta) = \sum_{i=1}^{m} \sum_{z^{(j)}} p(z^{(j)}|x^{(i)};\theta) \log \frac{p(x^{(i)}, z^{(j)};\theta)}{p(z^{(j)}|x^{(i)};\theta)}
$$

가 성립하고, 초기 파라미터 값을 설정하고 위 식을 최대화하는 M-step을 거침으로써 새로운 파라미터를 찾는다.

즉,

$$
\theta^{new} := \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(j)}} p(z^{(j)}|x^{(i)};\theta^{old}) \log \frac{p(x^{(i)}, z^{(j)};\theta)}{p(z^{(j)}|x^{(i)};\theta^{old})}
$$

이렇게 표현을 할 수가 있다. 여기서 헷갈리면 안되는 것이 θ^{old}는 우리가 지정한 초기 상수 값이고, θ는 최적화하고자 하는 모수다.

위 식은 아래 과정을 통해 간단하게 표현이 가능하다.

$$
\begin{aligned}
\theta^{new} := \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(j)}} p(z^{(j)}|x^{(i)};\theta^{old}) \log p(x^{(i)}, z^{(j)};\theta^{old}) - \sum_{i=1}^{m} \sum_{z^{(j)}} p(z^{(j)}|x^{(i)};\theta^{old}) \log p(z^{(j)}|x^{(i)};\theta^{old}) \\
\Leftrightarrow \Theta^{new} = \arg\max_{\Theta} \mathbf{E}[\log \mathbf{L}(\Theta;\mathbf{X},\mathbf{Z})|\mathbf{X},\Theta^{old}]
\end{aligned}
$$

여기서 두 번째 줄 term은 θ에 대해 독립이므로 아래 식이 도출된다.

기댓값 부분은 observed된 data X와 임의로 지정된 (초기에 설정한) 파라미터 Θ가 주어졌을 때 likelihood function의 기대값을 말한다. 이 기대값을 구할 때 z의 사후분포를 구하는 과정이 포함된다.

결론적으로 위 likelihood function의 기대값을 최대로 만드는 모수를 찾음으로써 한 번의 iteration이 끝난다.

정리하자면, EM 알고리즘은 정보를 최대한 활용하기 위해 latent variable을 포함해 ML 추정량을 구하는 iterative한 알고리즘이다. 기존의 MLE를 구하는 방법 (결합확률분포 구하고 미분값 = 0이 되는 지점 찾는 것)을 바로 적용하기 어려워 임의의 초깃값을 지정하고, Jensen 부등식을 이용해 결합확률분포의 하한을 구하여(E-step), 이를 최대화한다(M-step).

3. Convergence

EM 알고리즘과 같이 반복적인 알고리즘의 가장 중요한 점은 최적화하고자 하는 값이 수렴해야 한다는 것이다. EM 알고리즘은 파라미터가 업데이트 될수록 log-likelihood 값이 단조 증가하기 때문에 수렴하게 만드는 파라미터 역시 존재한다.

3. Examples

EM 알고리즘을 사용하는 예시 중 대표적인 mixture model에 대해 알아보자.

mixture model이란 여러 모델이 혼합된 형태로, introduction에서 예를 들었던 두 정규분포가 혼합된 경우도 mixture model이다.

아래 수식을 보면서 mixture model을 이해해보자

$$f_x(x) = \sum_j^m p_j \ \exp(-(x-\mu_j)^2/2\sigma_j^2)/\sqrt{2\pi\sigma_j^2}$$

위 mixture 모델은 m개의 정규분포가 혼합된 모델이다. 추정해야 하는 모수는 아래 세 개다.

- p_j : mixing coefficient, j번째 모델에서 표본이 등장할 확률을 결정
  $\sum_j p_j = 1, p_j \geq 0$
- μ_j & σ_j : j번째 정규분포의 평균과 분산

위 세 개의 모수를 찾기 위해 EM 알고리즘을 활용하게 되고, EM 알고리즘을 사용하기 위해 정규분포의 label을 가리키는 latent variable Z를 도입한다. Z는 다음이 성립한다.

$$P(Z=j) = p_j$$

또한,

$$f_{x|z}(x_i|z_i = j, \theta) = \frac{1}{\sqrt{2\pi\sigma_j}} e^{-(x_i-\mu_j)^2/2\sigma_j^2}$$

가 성립한다. θ는 모수 집합. (i.e. i번째 sample이 j번째 분포에서 나왔을 때의 확률 분포)

그렇다면 log-likelihood function은 다음과 같이 정의할 수 있다.

$$l(\theta; X, Z) = \sum_i \log p_{z_i} \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} e^{-(x_i-\mu_{z_i})^2/2\sigma_{z_i}^2}$$

위 식을 최대화하는 모수를 찾아야 하지만 latent variable의 존재로 그것은 어렵다. 따라서 EM 알고리즘을 사용해야 하고, 가장 먼저 log-likelihood function의 기대값을 구해야 한다.

1. E-step

$$E[l(\theta, X, Z)|X, \theta^{old}]$$
$$= \sum_i^n \sum_j^m P(Z_i = j|x_i, \theta^{old}) \log(f_{x|z}(x_i|z_i = j, \theta) P(Z_i = j|\theta))$$

여기서 베이즈의 정리를 이용해 $P(Z_i = j|x_i, \theta^{old})$를 정리할 수 있다.

$$P(Z_i = j|x_i, \theta^{old}) = \frac{P(Z_i = j, X_i = x_i|\theta^{old})}{P(X_i = x_i|\theta^{old})} = \frac{f_{x|z}(x_i|z_i = j, \theta) P(Z_i = j|\theta^{old})}{\sum_k^m f_{x|z}(x_i|z_i = k, \theta) P(Z_i = k|\theta^{old})}$$

위 식을 기대값을 구한 것에 집어 넣고 그 식을 최대화 하면 된다.

2. M-step

$$\theta^{new} = \arg\max_\theta E[l(\theta, X, Z)|X, \theta^{old}]$$

모수는 총 세 개이므로 각각 μ, σ^2, p로 편미분 해준 후, 미분값 = 0 을 만족시키는 모수로 업데이트함으로써 한 번의 EM 알고리즘이 동작한다.

참고 : EM_Algorithm.pdf (columbia.edu)

References

cs229-notes8.dvi (stanford.edu)

Lecture 14 - Expectation-Maximization Algorithms | Stanford CS229: Machine Learning (Autumn 2018) - YouTube